A Multi-agent Model Inspired on Soft-power

Francisco Coelho
fc@uevora.pt

July 23, 2015

Abstract

Soft-power is a relatively new player in the foreign affairs sciences. Instead of using coercion, characteristic of hard-power, it grows on the attraction of shared values. Multi-agent simulations are a tool of choice to study the social nature of soft-power, where realistic scenarios requiring heterogenous agents, acting rationally with respect to their own, individual, rewards can be explored.

The work described here is an experiment where deliberation results from a simple individual markov decision process that tries to incorporate elements inspired by soft-power. The resulting policies are then analyzed and tested in a virtual multi-agent simulation.

According to Robert Nye [6], in the context of foreign affairs, power is "*the ability to get the outcomes one wants*" and "*having the capabilities to affect the behavior of others to make those things happen.*" He argues that material, hard, resources are not enough to explain power and proposes another face of power, coined *soft-power*, that rests on the attraction of shared values. Soft-power players *choose* (co-opt) to work (cooperate) together, instead of being coerced to.

The role of individual values and the possibility of co-optation *versus* coercion can, arguably, be transported from the foreign affairs area to other social sciences, where agents interact to achieve common and individual goals. For example, motivated co-workers is a common goal for leaders; to build bridges from individual to institutional values (*e.g.* Google's motto, "*don't be evil*"[1]) can be described as exercises of soft-power.

Multi-agent models and computer simulations have became desirable, indeed unavoidable, tools to study social interactions. However, developing multi-agent models where the potential of individual deliberation is cripped is a dangerous game: Generalization of homogeneity in individuals is a common, old and actual,

[1] and not "*do no evil*", as is usually stated.

source of error and tragedy. The rush to understand the complexities of social interactions in simple terms overlooks the many individual deliberation processes that, in one hand, often drive the social dynamics out of local minima and, in the other hand, settle on diverse values and perception of the world states and dynamics.

Fortunately, individual deliberation can be formalized without bounds. Many authors argue that *rational decision*, based in *Bernoulli's* formula [3]

$$\alpha^*(x) \quad = \quad \arg\max_{\alpha} \sum_{x'} \mathbf{p}(x'|x,\alpha)\,\mathbf{u}(x') \tag{1}$$

is not adequate as a model of human deliberation because (a) humans are notoriously bad at assigning realistic values to transition probabilities $\mathbf{p}(x'|x,\alpha)$ and utilities $\mathbf{u}(x')$ and (b) the number of pairs (α, x') is too large to compute α^* in real-time. These statements are not in cause here: the volume of evidence supporting (a) is immense and (b) is directly observable.

However, such claims do not necessarily imply that Bernoulli's formula is wrong as an approximate model of human deliberation. Claim (a) merely implies that human deliberation doesn't use exact $\mathbf{p}()$ and $\mathbf{u}()$ functions. Moreover, at the population level, when individual evaluation errors tend to cancel out, there is some evidence that human behavior respects formula (1). For example, according to the general belief in economy sciences, (e.g. Grayson in [4]), the utility of money is proportional to the logarithm of the quantity. Under this supposition Bernoulli's formula predicts the well-known (to economy) behavior where wealth people tend to avoid risk while poor people tend to adopt it.

So, assuming that humans do indeed "have" some transition model, $\mathbf{p}()$, and utility, $\mathbf{u}()$, functions, possibly (perhaps certainly) erroneous ones, there is no reason to rule out Bernoulli's formula as a model of human deliberation: *the program is good but the input is bad.* Moreover, it seems that good approximations, and not complexity of the computation is the real issue: dogs catch objects in flight, although the trajectories are described by differential equations.

A model of human deliberation based in formula (1) might be adequate under the condition that at least one of "$\mathbf{p}()$", "$\mathbf{u}()$" or "$\arg\max_\alpha \sum_{x'}$" is not completely evaluated. In addition, the theory of rational decision, spawned from Bernoulli's formula, is a well studied mathematical framework rich in concepts, techniques and algorithms. In particular, Markov Decision Processes (MDP) express, in a compact language, the problem of finding an optimal policy given the transition model and (local) rewards.[2] The general structure of MDP problems is depicted in figure 1. For deliberation problems in particular, MDPs are extended to Dynamic Decision

[2] A *reward* measures the present state (and possibly the action and the successor state) whereas *utility* encompasses all the infinite horizons visible from the present state.

Networks (DDN), where the state x is factorized into features $x = \langle f_1, \ldots, f_n \rangle$. The major limitation of MDPs, and therefore of MDP based DDNs, is the assumption of *complete observability* of the current state. The realistic generalization to *partially observable* environments is formalized with Partially Observable Markov Decision Processes (POMDP) but finding exact solutions (*i.e.* approximating exactly the utility function) is known to be computationally intractable. However, recent results [10] provide algorithms able to approximate in real-time problems up to 10^{54} states.

In summary, soft-power, growing from shared values, accounts for a psychological face of power and rational decision might contribute with an adequate framework to model individual deliberation. These two ideas guide a process of exploration of models of rational deliberation incorporating elements from soft-power, the resulting policies from such models and multi-agent simulations, where individual agents follow such policies.

1 Deliberation and Soft-power

The task of defining a model of rational deliberation incorporating elements from soft-power can be divided into:

1. select an appropriate framework to express rational decision problems;
2. select deliberation features representing, at least partially, soft-power and
3. express the factors of step 2 in an instance of the framework of step 1.

In this section is formulated a general model of deliberation with soft-power features. It starts with a brief overview of MDP and proceeds to the statement of soft-power features and their interdependencies within the DDN framework.

1.1 Markov Decision Processes

Markov Decision Processes define the problem of finding an utility function, $\mathbf{u}(x)$, given a transition model, $\mathbf{p}(x'|x, \alpha)$, and a (local) reward function, $\mathbf{r}(x\,[, x', \alpha])$. The transition model expresses the markov condition: the (stochastic) state successor, x', depends solely on the current state, x, and on an (agent controlled) action, α; the reward function, $\mathbf{r}(x\,[, x', \alpha])$, depends on the current state and, possibly, on the action, α and the successor state, x'.

It turns out that Value-Iteration (VI) [8], a common algorithm to approximate the utility function, provides, *en passant*, an optimal policy, π^* — a function that associates states to actions ($\alpha^* = \pi^*(x)$) and that, by following such policy, the expected accumulated reward (*i.e.* the expected utility) is maximized.

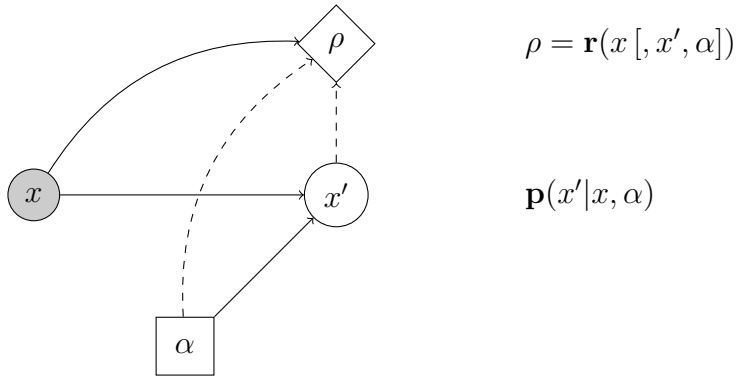

Figure 1: *Structure of a MDP problem.* Arrows denote dependency and nodes variables. The current state is x and the successor state is x', α is an (agent controlled) action and ρ is the (local) reward. The probability distribution $\mathbf{p}(x'|x, \alpha)$ is the transition model and $\mathbf{r}(x\,[, x', \alpha])$ the reward function. From the transition model and the reward function, the VI algorithm produces approximations to the utility function, $\mathbf{u}(x)$, and intermediary calculations promptly provide an approximation of the optimal policy, π^*.

The policy computed by the VI algorithm solves the deliberation problem of equation (1) since the task of selecting the optimal action, α^*, given the current state, x, is given by π^*:

$$\alpha^*(x) \quad \approx \quad \pi^*(x)\,. \tag{2}$$

Ideally, given a transition model and reward function, the VI algorithm is used once, before "run-time", and the resulting approximate policy, $\widetilde{\pi}$, defines the agent behavior. At "run-time", the agent merely (1) determines the current state, x, (hence the assumption of complete observability) and (2) uses the computed policy, $\widetilde{\pi}$, to choose the action associated to the current state, $\alpha = \widetilde{\pi}(x)$. In summary, MDP enables reactive agents with optimal behaviors. Of course, this procedure has too many weaknesses for real-world applications.

The paramount issue with MDPs is not the computational complexity of finding the optimal policy[3] but, instead, the methodological assumption that the agent always knows exactly the current environment's state. In POMDPs the assumption of *complete observability* is replaced by a *sensor model*. This class of problems was, until recently, computationally untreatable [7], even for limited applications. However, recent adaptations [9] of Monte-Carlo Tree-search algorithms [5], origi-

[3]Which is not insignificant, $O\left(|\mathbf{dom}(x)|^2 \times |\mathbf{dom}(a)|\right)$, but feasible.

nally applied to the game of Go [2] and Multiarmed bandit problems [1], handle large and complex POMDP problems.

1.2 Deliberation Features and Interdependencies

Despite its limitations, the formal and theoretical advantages of MDPs and DDNs provide a natural and rich framework to study deliberation with soft-power factors.

Perhaps the key core concept of soft-power is that of individual and social *values*. A distinctive assumption of soft-power is that shared values facilitate *co-optation* which, in turn, promotes *cooperation* that can be (grossly) translated as an effect in incoming and outgoing *resources*. Besides soft-power considerations, deliberation is directed to the realization of individual *goals*, that can be stated in terms of minimizing a "distance" from the (current, factual) *situation* to the (target, ideal) situation expressed by the values. The progress from the present situation towards the values depends on the available resources thus, indirectly, on cooperation and shared values. In short,

situation is the current, factual, "state-of-affairs";

values describe a target, ideal, situation;

goal is static in the sense that can be formulated as *minimizing the distance from the current situation to the values*;

cooperation results from shared values, provides and consumes resources;

resources are the means to achieve the goal;

So a simple model of deliberation can be defined as a DDN for states x, such that

$$x \;=\; \langle s, v, r, g, i, o \rangle \tag{3}$$

where, using DDN terminology, features $\langle s, v, r, g, i, o \rangle$ are described in table 1. The features, together with the agent action, interact to influence the value of the successor state (that, nevertheless, is described by a stochastic, markovian process). This evolution, from the current state to the successor state is the (agent's) transition model. The most general form would be

$$\mathbf{p}(x'|x, \alpha) \;=\; f\,(x', x, \alpha) \tag{4}$$

but specific (non-)interdependencies can be expressed stating the influences in each feature evolution. For example, the supposition that situation evolution depends

Table 1: *Description of features.* A decision state x is factorized into features $\langle s, v, r, g, i, o\rangle$. In this table only the **denotation** column is of consequence in computations.

name	**denotation**	**representing**	**soft-power?**
situation	s	current, factual, "state-of-affairs"	
values	v	target, ideal, "state-of-affairs"	yes
resources	r	means to progress towards values	
group-values	g	perception of social values	yes
income	i	(resource) contributions received	
outgoing	o	(resource) contributions given	

only on the current values and resources (*i.e.*, that doesn't depend on current group-values, income, outgoing or action), can be expressed by

$$\mathbf{p}(s'|s, v, r, g, i, o, \alpha) = \mathbf{p}(s'|s, v, r)$$

or, more succinctly, stating that situation evolution is given by

$$\mathbf{p}(s'|s, v, r)\,.$$

The specification of every feature dependencies refines the general equation (4). The deliberation model explored here has the following interactions:

situation: $\mathbf{p}(s'|s, v, r)$ situation evolution, $s, v, r \to s'$, depends on current values and resources;

values: $\mathbf{p}(v'|v, \alpha)$ values evolution, $v, \alpha \to v'$, depends only on the agent will, expressed by his action selection;

resources: $\mathbf{p}(r'|r, i, o, \alpha)$ resources evolution, $r, i, o, \alpha \to r'$, depends on the agent action, as well as on cooperation (income and outgoing resources);

group-values: $\mathbf{p}(g'|g, v, o)$ group-values evolution, $g, v, o \to g'$, can be affected by the volume of outgoing contributions;

income: $\mathbf{p}(i'|i, g, v)$ (cooperation) income, $i, g, v \to g'$, depends on the relation of individual values and group-values;

outgoing: $\mathbf{p}(o'|o, r, \alpha)$ (cooperation) outgoing, $o, r, \alpha \to o'$, results from the agent deliberation and available resources;

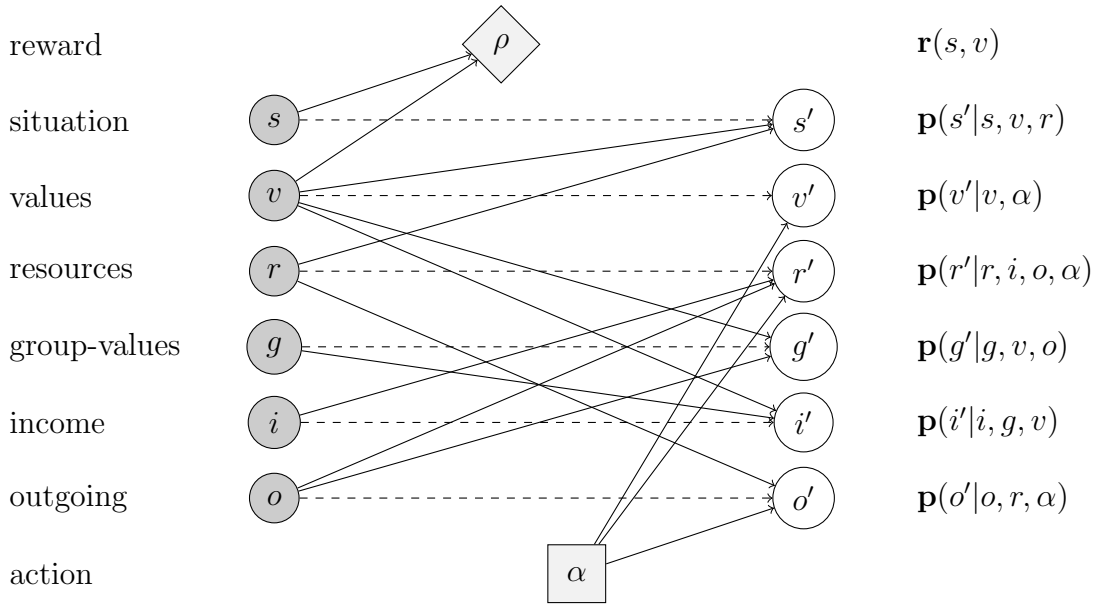

Figure 2: *Soft-power* DDN *model*. This diagram shows the interdependence of the features.

The resulting diagram for the transition model is depicted in figure 2. The agent's transition model results from joining all the interdependencies stated above into a single (probability) function of the deliberation state evolution:

$$\begin{aligned}\mathbf{p}(x'|x,\alpha) &= \mathbf{p}(s'|s,v,r)\,\mathbf{p}(v'|v,\alpha)\,\mathbf{p}(r'|r,i,o,\alpha)\times \\ &\quad\ \mathbf{p}(g'|g,v,o)\,\mathbf{p}(i'|i,g,v)\,\mathbf{p}(o'|o,r,\alpha) \end{aligned} \tag{5}$$

In addition, the agent "goal", defined in terms of a "distance" between the situation and the values, can be expressed by a reward function that depends only on the current situation and values:

$$\mathbf{r}(x) = \mathbf{r}(s,v) \tag{6}$$

A major limitation that could be pointed to this model is the lack of explicit individual-to-individual expression of soft-power. Indeed, the perception of soft-power seems amalgamated into a single variable, g. Formally, this shortcoming can be easily solved by expanding $g = \langle g_1, \ldots, g_N \rangle$ where each g_i is the individual perception of (another) agent's i individual values. The same expansion can be applied to incoming and outgoing resources. Furthermore, if different kinds of resources (*e.g.* human resources, stored goods, *etc.*) are to be accounted, so the variable r can be made multi-dimensional.

2 Policies Analysis

A general DDN model for deliberation with soft-power is defined in sub-section 1.2. The structure of the transition model is set in equation (5) and the reward function in equation (6). The VI algorithm applied to specific instances of such transition model and reward function returns an approximation of the optimal policy, $\widetilde{\pi}$.

One possible implementation of policies, for small scale problems, is using dictionaries, an indexed data structure. In this case state values are the indices and actions are the associated data. If P is a state-action dictionary then the action, α, to be executed in state x is given by

$$\alpha = \pi(x) \quad = \quad P[x].$$

Also the implementation of the transition model and the reward function can be based on dictionaries. For example, the transition of the situation variable can be defined by

$$\mathbf{p}(s'|s,v,r) \quad = \quad S[s',s,v,r].$$

Of course, the explicit expression of such tables is a tedious work. Denoting by $\mathbf{dom}(var)$ the domain of $var \in \{\alpha, s, v, r, g, i, o\}$ and $|X|$ the cardinality of set X, at each deliberation step, the agent can choose one of $|\mathbf{dom}(\alpha)|$ actions. Also, the unprocessed input of the VI algorithm is a list of seven dictionaries, for the transition model and reward function, with sizes easy to find: for example, a table representing $\mathbf{p}(s'|s,v,r)$ has $|\mathbf{dom}(s)|^2 \times |\mathbf{dom}(v)| \times |\mathbf{dom}(r)|$ keys.

2.1 A Minimal Case

If each feature can take one of two different values and the agent can choose one of two actions, there are $64 = 2^6$ deliberation states and the model requires $104 = 4 \times 2^4 + 2^3 + 2^5$ parameters, plus 4 for the reward function. In this case, since there are 64 deliberation states, the policy output of the VI algorithm is a dictionary with 64 keys.

Instead of having hand-written the $104 + 4$ parameters, the transition model and the reward function can both be generated by a set of adequate rules. Supposing that the transition probabilities of the various variables result from direct observations (entailing *a posteriori*, empiric, probabilities), these probabilities can be computed from a complete set of rules that define the number of observations of each possible scenario. For example, to define $\mathbf{p}(s'|s,v,r)$, the rule

```
s-obs(s', s, v, r) := r == 1 and s' == v ? 3 : 1
```

states that, in all possible transitions $x = (s, v, r, g, i, o) \rightarrow x' = (s', v', r', g', i', o')$ each case that matches the condition $r = 1 \wedge s' = v$ has 3 observations and each one of the remaining cases has one observation. This observation rule creates a bias towards the cases $r = 1, s' = v$ in the transition probability of $s, v, r \rightarrow s'$ that can be translated as "*when resources are high,* $r = 1$*, the cases where the next situation matches the values,* $s' = v$*, are more probable*".

The complete transition model, with formal rules detailed in Appendix A, can be defined for each feature:

situation: *It is more probable that the next situation equals the values when resources are high.*

values: *Action "0" tends to keep the current values.*

resources: *Action "1" tends to increase resources and action "0" to decrease.*

group-values: *The group-values approach individual values but faster if outgoing-cooperation is high.*

income: *High income is more probable when values match group-values.*

outgoing: *High outgoing is more probable when resources are high and action is "0".*

The reward function valorizes states where the situation equals the values, $s = v$, and gives a penalty otherwise. The reward function is defined by

reward: *states where the situation equals the values,* $s = v$*, have a positive reward and other states have a negative reward.*

```
reward(s, v) := s == v ? +1 : -1
```

These rules entail a explicit transition model and reward function. Then, with these inputs, the VI algorithm returns an approximation of the optimal policy, assigning an action for each one of the 64 possible deliberation states. Appendix B shows the output policy. It turns out that, by direct examination, such policy can be simplified to the following rule:

```
action(s, v, r, g, i, o) :=
    r == 1 ? 0
           : ( g == 0 ? 1
                      : v == o ? 0 : 1
             )
```

An immediate, attractive, interpretation of this rule is that *wealth agents can insist on current values, regardless of all other variables.* Also, neither the current situation, s, nor income, i, are relevant to deliberation — a clear sign of the shortcomings of this particular instance as a candidate model of real-world soft-power cases.

3 Multi-agent Simulation

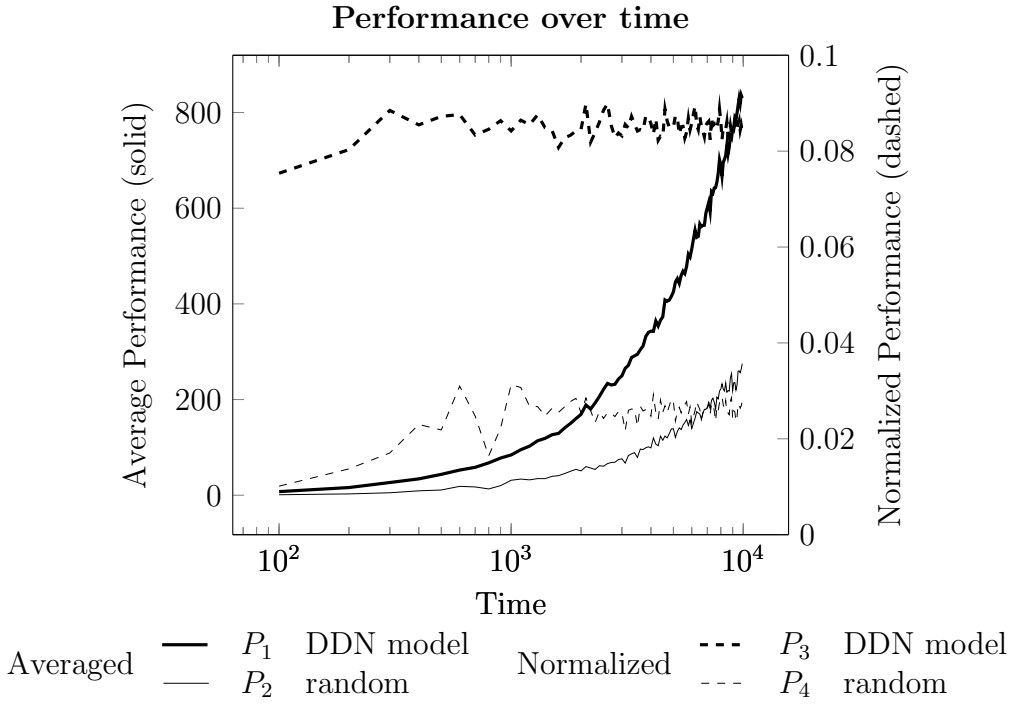

Figure 3: *Average and normalized performance over time.* This double plot shows data from a set of 30 simulations where average agent performance —the accumulated reward divided by the "agent age"— is recorded after some time (in the horizontal axis). The profile of agents applying the policy computed by the DDN model (P_1 and P_3) is represented with thick lines versus random agents (P_2 and P_4), with thin lines. Solid lines (P_1 and P_2) show the average performance and the dashed lines (P_3 and P_4) normalize these values by the theoretic maximum performance value.

The previous section ended with an interpretation of the policy returned from the VI algorithm, given a certain transition model and reward function. An im-

mediate quest from that point on is to find out how does that particular policy affects the agent performance.

One possibility to study the policy performance is through simulation. But the environment evolution doesn't need to comply with rules of the transition model of the agent: agents can be mistaken about the environment evolution. More specifically, the rules that guided the environment simulation, described in Appendix C, do not match the transition model from subsection 2.1.

The simulation procedure places an agent implementing some policy in a virtual environment and the values of successive rewards are recorded. The performance of a reference agent, with random action selection, is also recorded. The simulation results are summarized in figure 3.

4 Discussion

4.1 What Was Done

This work starts properly, in section 1.2, by proposing a basic model of rational deliberation with (intended) soft-power features (situation, values, resources, *etc*), formalized within the DDN framework, that encompasses MDPs has a "rationality" engine. Soft-power is directly expressed in this model by two features: *values* and *group-values*. Further influence of these features transit into the full deliberation state.

In subsection 2.1 is defined a minimal case of that model, using the rules in appendix A to generate a transition model and reward function. These are then processed by the VI algorithm that outputs an approximation of the optimal policy (for the minimial case), shown in appendix B.

Such policy defines the behavior of agents in a multi-agent simulation, where the performance of such agents is related, in section 3, to the performance of agents behaving randomly. The environment evolution in this simulation is guided by the stochastic rules in appendix C.

4.2 What Remains To Be Done

The features and interdependencies of the proposed model lack any justification. Therefore, it is mandatory to ground the structure of the model with real cases and previous studies. Until such relations can be made, this work remains little more than an abstract exercise. Candidates to provide such links include, but are not limited to, the social sciences in general (psychology, economy, international affairs, *etc.*).

Soft-power can not be reduced, as is done here, to group and individual values features. Only the aspect of cooperation, by itself, provides a vast field of research. Furthermore, there is clear need to frame any reasonable model of individual deliberation within a sound psychologic background.

On a different line of future development, the assumption of complete observability, to support a MDP based model, is not realistic. This can be fixed by moving to a POMDP framework, where recent results provide hope for feasible computational approaches.

A Transition Model Rules

These rules define the agent's belief about the transition model of the environment and are used by the agent to compute an (believed to be) optimal policy.

situation: *It is more probable that the next situation equals the values when resources are high.*

```
s-obs(s', s, v, r) := r == 1 and s' == v ? 3 : 1
```

values: *Action "0" tends to keep the current values.*

```
v-obs(v', v, alpha) :=
    alpha == 0 ? (v' == v ? 3 : 1)
               : 1
```

resources: *Action "1" tends to increase resources and action "0" to decrease.*

```
r-obs(r', r, i, o, alpha) :=
    alpha == 1 ? (r' == 1 ? 2 : 1)
               : (r' == 1 ? 1 : 2)
```

group-values: *The group-values approach individual values but faster if outgoing-cooperation is high.*

```
g-obs(g', g, v, o) :=
    g' == v ? 2
            : (o == 1 ? 1 : 2)
```

income: *High income is more probable when values match group-values.*

```
i-obs(i', i, v, g) :=
    i' == 1 and v == g ? 2
                       : 1
```

outgoing: *High outgoing is more probable when resources are high and action is "0".*

```
o-obs(o', o, r, alpha) :=
    alpha == 0 and r == 1 and o' == 1 ? 2
                                      : 1
```

B Approximate Optimal Policy

This table lists the policy computed by the VI algorithm given the transition model derived from the rules in appendix A and reward in subsection 2.1.

s	v	r	g	i	o	α	s	v	r	g	i	o	α	s	v	r	g	i	o	α
0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	1
0	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	1	0	1	1
0	0	0	1	1	0	0	0	0	0	1	1	1	1	0	0	1	0	0	0	0
0	0	1	0	0	1	0	0	0	1	0	1	0	0	0	0	1	0	1	1	0
0	0	1	1	0	0	0	0	0	1	1	0	1	0	0	0	1	1	1	0	0
0	0	1	1	1	1	0	0	1	0	0	0	0	1	0	1	0	0	0	1	1
0	1	0	0	1	0	1	0	1	0	0	1	1	1	0	1	0	1	0	0	1
0	1	0	1	0	1	0	0	1	0	1	1	0	1	0	1	0	1	1	1	0
0	1	1	0	0	0	0	0	1	1	0	0	1	0	0	1	1	0	1	0	0
0	1	1	0	1	1	0	0	1	1	1	0	0	0	0	1	1	1	0	1	0
0	1	1	1	1	0	0	0	1	1	1	1	1	0	1	0	0	0	0	0	1
1	0	0	0	0	1	1	1	0	0	0	1	0	1	1	0	0	0	1	1	1
1	0	0	1	0	0	0	1	0	0	1	0	1	1	1	0	0	1	1	0	0
1	0	0	1	1	1	1	1	0	1	0	0	0	0	1	0	1	0	0	1	0
1	0	1	0	1	0	0	1	0	1	0	1	1	0	1	0	1	1	0	0	0
1	0	1	1	0	1	0	1	0	1	1	1	0	0	1	0	1	1	1	1	0
1	1	0	0	0	0	1	1	1	0	0	0	1	1	1	1	0	0	1	0	1
1	1	0	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	0	1	0
1	1	0	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0
1	1	1	0	0	1	0	1	1	1	0	1	0	0	1	1	1	0	1	1	0
1	1	1	1	0	0	0	1	1	1	1	0	1	0	1	1	1	1	1	0	0
1	1	1	1	1	1	0														

C Environment Evolution Rules

These rules define the environment state evolution. Reading `prob(p)` as "true with probability `p`", the successor state, x', generated by the environment, when the agent is in state x and chooses action α is defined by the following rules:

if action $\alpha = 0$:

```
s' := r == 1 ? ( prob(0.6) ? v : s )
             : ( prob(0.4) ? v : s )
v' := prob(0.9) ? v
                : 1 - v
r' := r == 1 ? ( prob(0.4) ? r : 1 - r )
             : r
g' := g == v ? ( prob(0.9) ? v : 1 - v )
             : ( prob(0.8) ? g : 1 - g )
i' := g == v ? ( prob(0.6) ? 1 : 0 )
             : ( prob(0.4) ? 1 : 0 )
o' := r == 1 ? ( prob(0.3) ? 1 : 0 )
             : ( prob(0.1) ? 1 : 0 )
```

if action $\alpha = 1$:

```
s' := r == 1 ? ( prob(0.6) ? v : s )
             : ( prob(0.4) ? v : s )
v' := prob(0.2) ? v : 1 - v
r' := prob(0.6) ? 1 : 0
g' := prob(0.1) ? 1 - g : g
i' := g == v ? ( prob(0.6) ? 1 : 0 )
             : ( prob(0.4) ? 1 : 0 )
o' := r == 1 ? ( prob(0.3) ? 1 : 0 )
             : ( prob(0.1) ? 1 : 0 )
```

References

[1] Peter Auer and N Cesa-Bianchi. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine learning*, pages 235–256, 2002.

[2] Rémi Coulom. Efficient selectivity and backup operators in Monte-Carlo tree search. *Computers and Games*, pages 72–83, 2007.

[3] G. Cramer. 1728, letter from cramer to nicholas bernoulli. translated into english by louise sommer in bernoulli, daniel, 1954," exposition of a new theory on the measurement of risk,". *Econometrica*, 22:23–36.

[4] C. J. Grayson. Decisions under uncertainty: Drilling decisions by oil and gas operators. Technical report, Division of Research, Harvard Business School, Boston, 1960.

[5] Levente Kocsis and C. Szepesvári. Bandit based monte-carlo planning. *Machine Learning: ECML 2006*, pages 282–293, 2006.

[6] J.S. Nye. *Soft power: The means to success in world politics*. Public affairs, 2004.

[7] Joelle Pineau. Anytime Point-Based Approximations for Large POMDPs. *Machine Learning*, 27:335–380, 2006.

[8] D.L. Poole and A.K. Mackworth. *Artificial Intelligence - Foundations of Computational Agents*. Cambridge University Press, 2010.

[9] Stéphane Ross, Joelle Pineau, Sébastien Paquet, and Brahim Chaib-Draa. Online Planning Algorithms for POMDPs. *The journal of artificial intelligence research*, 32(2):663–704, July 2008.

[10] David Silver and Joel Veness. Monte-Carlo planning in large POMDPs. *Advances in Neural Information Processing Systems (NIPS)*, pages 1–9, 2010.